Home Assignment 1

ECE 602 - Introduction to Optimization

Due: March 6, 2020

Exercise 1

Explain which of the following sets and functions are convex and which are not. Explain your answers.

(a) The sublevel set of a convex function f, i.e., $S_\alpha = \{x \in \mathbb{R}^n \mid f(x) \le \alpha\}$.

(b) The set of points whose distance to a does not exceed a fixed fraction θ of the distance to b, i.e., the set $\{x \mid ||x-a||_2 \le \theta ||x-b||_2\}$. You can assume $a \neq b$ and $0 \le \theta \le 1$.

(c) Suppose $A \in \mathcal{S}^n, b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. The set $C = \{x \in \mathbb{R}^n \mid x^T A x + b^T x + c \le 0\}$ if A is negative definite matrice. **Hint**: A set is convex if and only if its intersection with an arbitrary line is convex.

(d) $f(x) = \frac{||Ax+b||_2^2}{c^T x + d}$ on $\{x \in \mathbb{R}^n \mid c^T x + d > 0\}$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $d \in \mathbb{R}$.

(e) $f(x) = g(h(x))$ where $h : \mathbb{R}^n \to \mathbb{R}$ is convex, while $g : \mathbb{R} \to \mathbb{R}$ is convex and monotonically increasing.

(f) $f(x) = (\sum_{i=1}^n x_i^p)^{\frac{1}{p}}$ with **dom** $f = \mathbb{R}^n_{++}$ and $p < 1, p \neq 0$. **Hint** : You may want to use the Cauchy-Schwartz inequality $|a^T b| \le ||a||_2 ||b||_2$.

Exercise 2

Let D_c be an $n \times n$ matrix defined as

$$D_c = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix}$$

and let D_r be an $m \times m$ matrix defined in a similar way. Then, given some $X \in \mathbb{R}^{n\times m}$, one can define its (discrete) *total variation* (TV) as

$$f(X) := \|X\|_{\mathrm{TV}} = \mathrm{Trace}\Big(\mathbf{1}_{n\times m}^T \sqrt{(D_c X)^2 + (X D_r^T)^2}\Big),$$

where the power and square root functions are applied coordinate-wise (i.e., *diagonally*). Note that $D_c X$ computes the column-wise differences of X, while $X D_r^T$ computes the differences along its rows.

(a) Is $f(X)$ a norm? Please explain.

(b) Is $f(X)$ a convex function? What about strict convexity?

(c) Using the *external* definition of derivative/gradient, derive a closed-form expression for $\nabla f(X)$.

Exercise 3

Suppose we are given a set of N (explanatory) variables $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$, and their corresponding observations $\{y_i\}_{i=1}^N$, with $y_i \in \mathbb{R}$, for all $i = 1, 2, \ldots, N$. For the sake of convenience, the available data can be arranged into a list of ordered pairs $S_N := \{x_i, y_i\}_{i=1}^N$, which will be referred below as a *training set.*

In general, the dependency between x and y is unknown, and our objective is to approximate it by means of a parametric function

$$f(\cdot \mid \Theta) : \mathbb{R}^d \to \mathbb{R} : x \mapsto y = f(x \mid \Theta),$$

where $\Theta \in \mathbb{R}^D$ is a vector of adjustable parameters. In particular, we are interested to find an *optimal* Θ^* that minimizes the *mean-squared error* (MSE), namely

$$\Theta^* = \arg\min_{\Theta} \big\{E(\Theta)\big\}, \quad E(\Theta) = \frac{1}{N}\sum_{i=1}^N |f(x_i \mid \Theta) - y_i|^2.$$

To render the problem practicable, we are going to restrict f to a particular class of functions. To this end, for $l = 1, 2, \ldots, L$, let us first define $f_l : \mathbb{R}^{M_{l-1}} \to \mathbb{R}^{M_l}$ as

$$f_l(u \mid W_l, b_l) = \varphi_l(W_l u + b_l), \qquad u \in \mathbb{R}^{M_{l-1}},$$

with $W_l \in \mathbb{R}^{M_l \times M_{l-1}}$, $b_l \in \mathbb{R}^{M_l}$ and with functions $\varphi_l : \mathbb{R} \to \mathbb{R}$ applied in a component-wise manner (i.e., diagonally). Moreover, we require $M_0 = d$ and $M_L = 1$ and assume that $\varphi_1(\tau) = \varphi_2(\tau) = \ldots = \varphi_{L-1}(\tau) \equiv \max(\tau, 0)$, while φ_L is an identity.

Subsequently, we define f by a sequence of recursive computations according to

$$u_l = f_l(u_{l-1} \mid W_l, b_l),$$

with $u_0 = x$ and $u_L = y$. Note that the resulting function depends on the entire set of parameters $W_1, b_1, W_2, b_2, \ldots, W_L, b_L$, which can be collected into one long vector Θ. Note that $f(x \mid \Theta)$ can also be viewed as a composition of functions f_l, *viz.*

$$f(x \mid \Theta) = f_L(\cdot \mid W_L, b_L) \circ f_{L-1}(\cdot \mid W_{L-1}, b_{L-1}) \circ \ldots f_2(\cdot \mid W_2, b_2) \circ f_1(x \mid W_1, b_1),$$

which is a popular mathematical model for an *Artificial Neural Network.*

(a) Is $E(\Theta)$ a convex function?

(b) What is the total number of model parameters D?

(c) Derive the gradient $\nabla E(\Theta)$ based on the definition of external derivatives.

(d) Write a MATLAB function which, for a given Θ, will compute the value of $\nabla E(\Theta)$ *by means of back-propagation.*

(e) Let $f_0 : \mathbb{R}^2 \to \mathbb{R}$ be defined as

$$f_0(x) = \frac{\sin(\|x\|_2)}{\|x\|_2}, \text{ with } \operatorname{dom} f_0 = \mathcal{B}_\infty(0, 4\pi).$$

For $N = 5000$, draw samples x_i uniformly from $\operatorname{dom} f_0$ and set $y_i = f_0(x_i)$, for all $1 \leq i \leq N$. Also, set $L = 6$ and $M = [2, 5, 9, 5, 3, 1]$.

Using the above parameters and data, compute Θ^* by means of

- Gradient Descent Method both with a fixed step size and using backtracking;
- Conjugate Gradient Method both with a fixed step size and using backtracking.

At each iteration, monitor the values of $E(\Theta)$ and of $\|\nabla E(\Theta)\|$. Are they converging monotonously?

(f) Plot the values of the resulting approximation $f(x \mid \Theta^*)$ over a uniform square grid and compare these values to those of f_0. Do they look similar? If not, try to explain why.

(g) How would you change the network design to improve the quality of the approximation. If $f_0(x)$ is replaced by a different function $f_0(x) = \|x\|_\infty$, does the quality of your approximation change? If yes, why?